Genome-Wide Analysis of Repetitive Elements in Papaya

Niranjan Nagarajan · Rafael Navajas-Pérez · Mihai Pop · Maqsudul Alam · Ray Ming · Andrew H. Paterson · Steven L. Salzberg

Received: 7 August 2008 / Accepted: 8 September 2008

Abstract Papaya (*Carica papaya* L.) is an important fruit crop cultivated in tropical and subtropical regions worldwide. A first draft of its genome sequence has been recently released. Together with *Arabidopsis*, rice, poplar, grapevine and other genomes in the pipeline, it represents a good opportunity to gain insight into the organization of plant genomes. Here we report a detailed analysis of repetitive elements in the papaya genome, including transposable elements (TEs), tandemly-arrayed sequences, and high copy number genes. These repetitive sequences account for ~56% of the papaya genome with TEs being the most abundant at 52%, tandem repeats at 1.3% and high copy number genes at 3%. Most common types of TEs are represented in the papaya genome with retrotransposons being the dominant class, accounting for 40% of the genome. The most prevalent retrotransposons are Ty3-gypsy (27.8%) and Ty1-copia (5.5%). Among the tandem repeats, microsatellites are the most abundant in number, but represent only 0.19% of the genome. Minisatellites and satellites are less abundant, but represent 0.68% and 0.43% of the genome, respectively, due to greater repeat length. Despite an overall smaller gene repertoire in papaya than many other angiosperms, a significant fraction of genes (>2%) are present in large gene families with copy number greater than 20. This repeat database clarified a major part of the papaya genome organization and partly explained the lower gene repertoire in papaya than in *Arabidopsis*.

Keywords Papaya genome · Transposable elements · Tandem repeats · Satellite DNA · High copy-number genes · Repeatome

Introduction

Much of the nuclear genome of most angiosperms is composed of different repetitive DNA elements. This was first stated by Thomas [47], who coined the term C-value paradox to describe the observation that genome size does not always correlate with structural complexity, and that variations in DNA content are mainly due to the accumulation of such repetitive sequences. Plant genomes have acquired a variety of repeat elements that account for up to 97% of nuclear DNA [14, 31]. For practical purposes, repetitive sequences can be divided into three main classes. (1) Transposable elements (TEs), which are the best-defined category and constitute the most abundant component of many genomes, ranging from 40% to 80% [4]. TEs can be further divided into RNA-mediated class I retrotransposons and DNA-mediated class II transposons. The most common TEs in plants are LTR retrotransposons [3]. In contrast, transposition of non-LTR retrotransposons is rarely observed in plants, suggesting that most of them are

Nagarajan and Navajas-Pérez contributed equally to this work.

N. Nagarajan (✉) · M. Pop · S. L. Salzberg
Center for Bioinformatics and Computational Biology,
University of Maryland,
College Park, MD 20742, USA
e-mail: niranjan@umiacs.umd.edu

R. Navajas-Pérez · A. H. Paterson
Plant Genome Mapping Laboratory, University of Georgia,
Athens, GA 30602, USA

M. Alam
Department of Microbiology, University of Hawaii,
Honolulu, HI 96822, USA

R. Ming
Department of Plant Biology,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801, USA

inactive and/or under regulation of the host genome [7]. (2) Tandem repeats, where individual copies are arranged adjacent to each other forming tandem arrays of the monomeric unit, appear preferentially in the centromeric, telomeric, and subtelomeric regions of many eukaryotes, comprising hundreds or thousands of repeats [19, 49]. These repeats also appear at interspersed positions and in low-recombining regions, such as sex chromosomes or B chromosomes [6, 33], with their function, if any, still unclear. These type of sequences can account for a large portion of genomic DNA in some cases [15, 38]. Finally, (3) high copy number genes, such as ribosomal or histone genes, are also an important part of the repeatome.

Except for high copy number genes, repetitive elements have often been considered junk DNA with no function [35]. However, recent studies suggest that they may play an important role as drivers of genome evolution in several regards, such as response to environmental cues [44], determination of continuous phenotypic characters [27] and gene regulation [48].

The study of repetitive sequence elements is essential to our understanding of the nature and consequences of genome size variation between different species, and for studying the large-scale organization and evolution of plant genomes. As a result, different databases and methods devoted to the analysis of this type of DNA have been recently developed (see for example [5, 10, 18, 25, 28, 32, 50, 51]).

Here we describe CPR-DB, a database of papaya genome repeats, in an effort to shed light on papaya genome organization and specifically on the role of repetitive elements. It should additionally be a valuable resource for the study of angiosperm evolution by facilitating the rapid identification and characterization of repetitive elements in other related plants.

Results

In order to create a papaya repeat database (CPR-DB), a papaya female genome sequence [30] was mined for repetitive elements. CPR-DB is divided into three main categories: transposable elements (TEs), tandem repeats and high copy number genes.

Transposable Elements

TEs are abundant in the papaya genome, with more than 43.4% of the genome (Table 1) being homologous to identifiable TEs. An additional 8.5% of the genome is covered by repetitive sequences that are currently unannotated but are likely to be novel TEs, based on their high copy number and similarity to other TEs. Thus, about 52% of the papaya genome is composed of TEs.

Most common types of TEs are represented in the papaya genome (among more than 600 types in Repbase [18]) with the dominant class being retrotransposons (40% of the genome) and the most abundant types being *Ty3-gypsy* (27.8%) and *Ty1-copia* (5.5%) retrotransposons and CACTA-like DNA transposons (0.01%). An interesting feature of the papaya genome seems to be the relatively low abundance of known transposons (0.20%) compared to other plant genomes. Some of this discrepancy could be accounted for by the presence of several papaya-specific transposon families that have yet to be annotated.

Table 1 Summary of TE content of papaya contigs

Class	Element	Length occupied	Percent of total contig length
I (Retrotransposons)	*Ty3-gypsy*	77.3 Mbp (76.8 Mbp)	27.8 (27.6)
	Ty1-copia	15.3 Mbp (14.1 Mbp)	5.5 (5.1)
	LINE	3.0 Mbp (2.7 Mbp)	1.1 (0.96)
	SINE	1.1 kbp	<0.01
	Other	23.6 Mbp (22.0 Mbp)	8.4 (7.9)
II (Transposons)	CACTA, En/Spm	40.2 kbp	0.01
	Micron	9.7 kbp	<0.01
	MITE	7.6 kbp	<0.01
	MuDR-IS905	7.6 kbp	<0.01
	Other	497.8.0 kbp (469.1 kbp)	0.18 (0.17)
Unclassified	Unknown	598.2 kbp (70.7 kbp)	0.22 (0.03)
Unannotated	Unknown	23.7 Mbp (23.7 Mbp)	8.5 (8.5)
Total	Misc.	144.1 Mbp (140.8 Mbp)	51.9 (50.4)

Papaya contigs were compared to known repeat elements in Repbase, the TIGR plant repeat database and a collection of papaya specific TE sequences, using RepeatMasker (with a conservative cutoff score of 350). The numbers in parentheses are those for the papaya specific repeat sequences

The annotation of TEs, in general, is a difficult task as the lack of selective constraints on the element can make it unidentifiable by homology in an evolutionarily short time-span. To annotate the papaya genome, therefore, we would ideally need a library of TE families that are papaya-specific or from a closely related species. Due to the absence of such a dataset, we constructed our own library, using de novo repeat finders to identify repeat families in the papaya genome and curating and annotating those that correspond to TEs (see "Methods"). We hope that our curated database of 889 papaya TE families can now serve as a resource in the annotation of other plant genomes.

For annotating papaya contigs, we used sequences in Repbase [18] and TIGR plant repeats (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats) in addition to the papaya TE families to do homology searches with RepeatMasker (http://www.repeatmasker.org). Not surprisingly, the transposon annotation of papaya contigs is dominated by the set of papaya-specific TE families (Table 1). In fact, when these sequences are excluded from the analysis only 14% of the genome is annotated as TEs (Table 2), thus demonstrating their utility in the analysis.

Typically, the matches to TEs in the genome tend to be inactive fossil repeats that have diverged from the consensus sequence. However, for several families of papaya-specific repeats (many of them annotated as *Ty3-gypsy* elements) we found dozens of nearly perfect copies, suggesting the possibility of some recent activity. Interestingly, some of the annotated papaya-specific repeat families also match EST sequences from other plant species.

To investigate the evolutionary history of TE elements in the papaya genome we reconstructed evolutionary trees for three known plant TE sequences (Ty1-Copia: RN107_I, FRSgTERT00100296 and Ty3-Gypsy: ATGP5A_I) with many good matches in the papaya genome (see "Methods"). As can be seen in Figs. 1 and 2, sequences from the same plant species typically form a distinct clade with high confidence and the overall topology of the high confidence edges matches the known plant phylogeny. In addition, the various sequences exhibit different patterns of conservation in the five published plant genomes. For example, Fig. 1 demonstrates an interesting feature of matches to the Ty3-Gypsy retrotransposon ATGP5A_I: it has two distinct regions of conservation. While the poplar genome has many good matches to bases 1,200–1,500 and none to bases 3,700–4,100 in ATGP5A_I, the reverse is true for the rice genome. Similarly, in Fig. 2, while both the Ty1-Copia sequences have one region of conservation the set of species that match them is markedly different. Among the two species that are common, sequences from the grape genome are more conserved than those from the papaya genome, suggesting a more recent introduction of these sequences into the grape genome or slower evolution in the grape genome. Interestingly, in both Figs. 1 and 2, sequences from the papaya genome tend to cluster more closely than expected with sequences from the rice genome.

Tandem Repeats

The papaya genome was also scanned for tandem repetitive elements—between 1 and 2,000 bp—using the Tandem Repeats Finder software [5]. A total of 414,681 repeats were characterized in 57,360 loci (spanning a total of 4.8 Mbps, representing 1.3% of the total genome size). The analysis revealed an average repetitive-unit length of 79 bp and a copy number average of 7.23 (ranging from 1.8 to 969.3 copies). The average AT content was 72%, slightly higher than the average AT content of the genome (65%).

Tandem repeats were classified into three classes: microsatellites (1–6 bp), minisatellites (7–100 bp) and satellites (>100 bp). Table 3 summarizes the sizes and abundance of tandem repeats. In terms of physical quantity

Table 2 Summary of plant repeat element content of papaya contigs

Class	Element	No.	Length occupied	Percent of total contig length
I (Retrotransposons)	*Ty3-gypsy*	27,964	20.8 Mbp	7.51
	Ty1-copia	13,816	8.1 Mbp	2.93
	LINE	1,367	0.3 Mbp	0.11
	SINE	37	3.2 kbp	<0.01
	Other	15,599	4.4 Mbp	1.56
II (Transposons)	CACTA, En/Spm	4,591	359 kbp	0.13
	MuDR-IS905	675	53.6 kbp	0.02
	Tc1-IS630-Pogo	380	20.5 kbp	0.01
	Other	5,702	437 kbp	0.15
Unclassified	Unknown	43,638	6.4 Mbp	2.30
Total	Misc.	113,789	40.8 Mbp	14.72

Papaya contigs were compared only to known repeat elements in Repbase and the TIGR plant repeat database using RepeatMasker (with the less conservative default parameters)

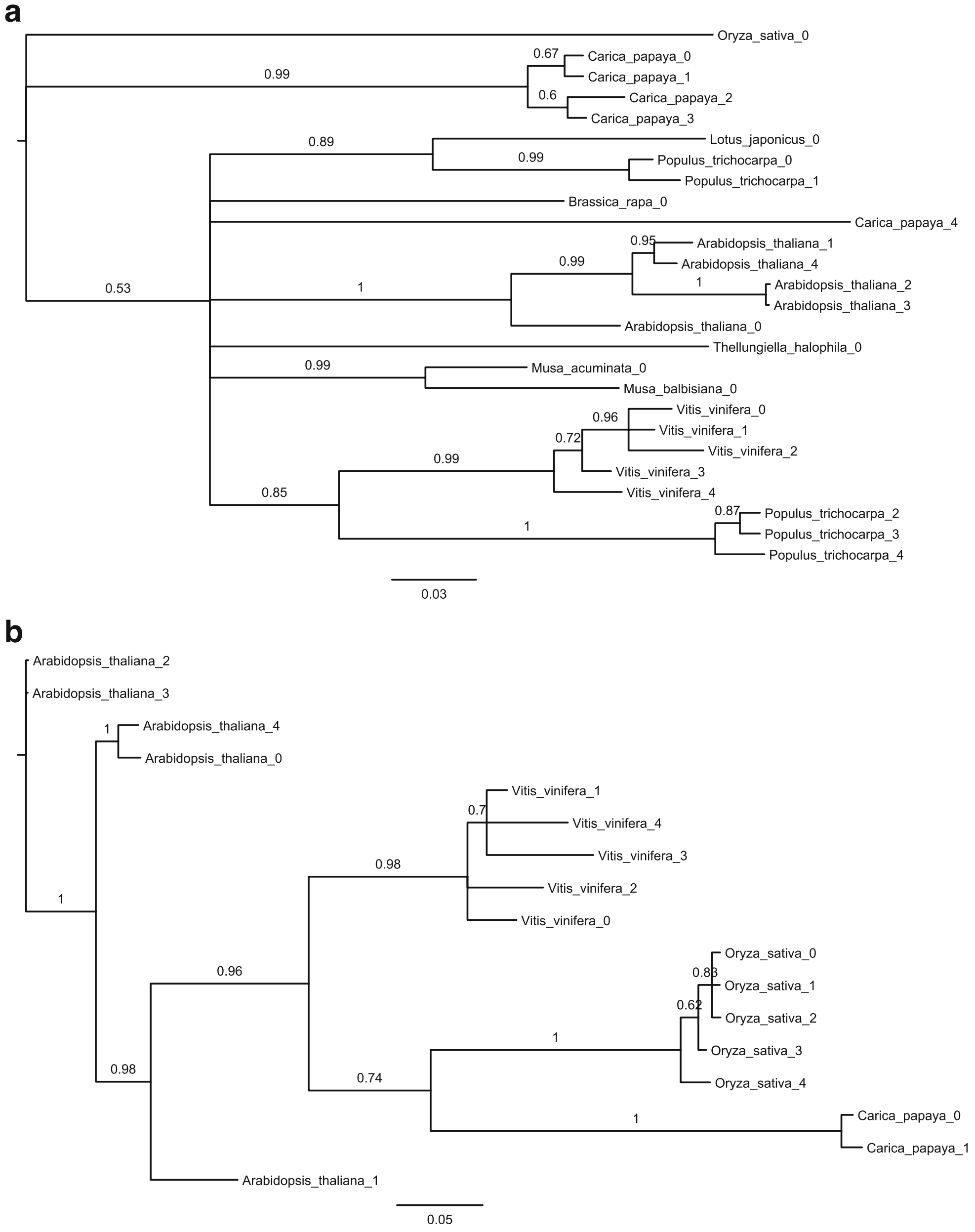
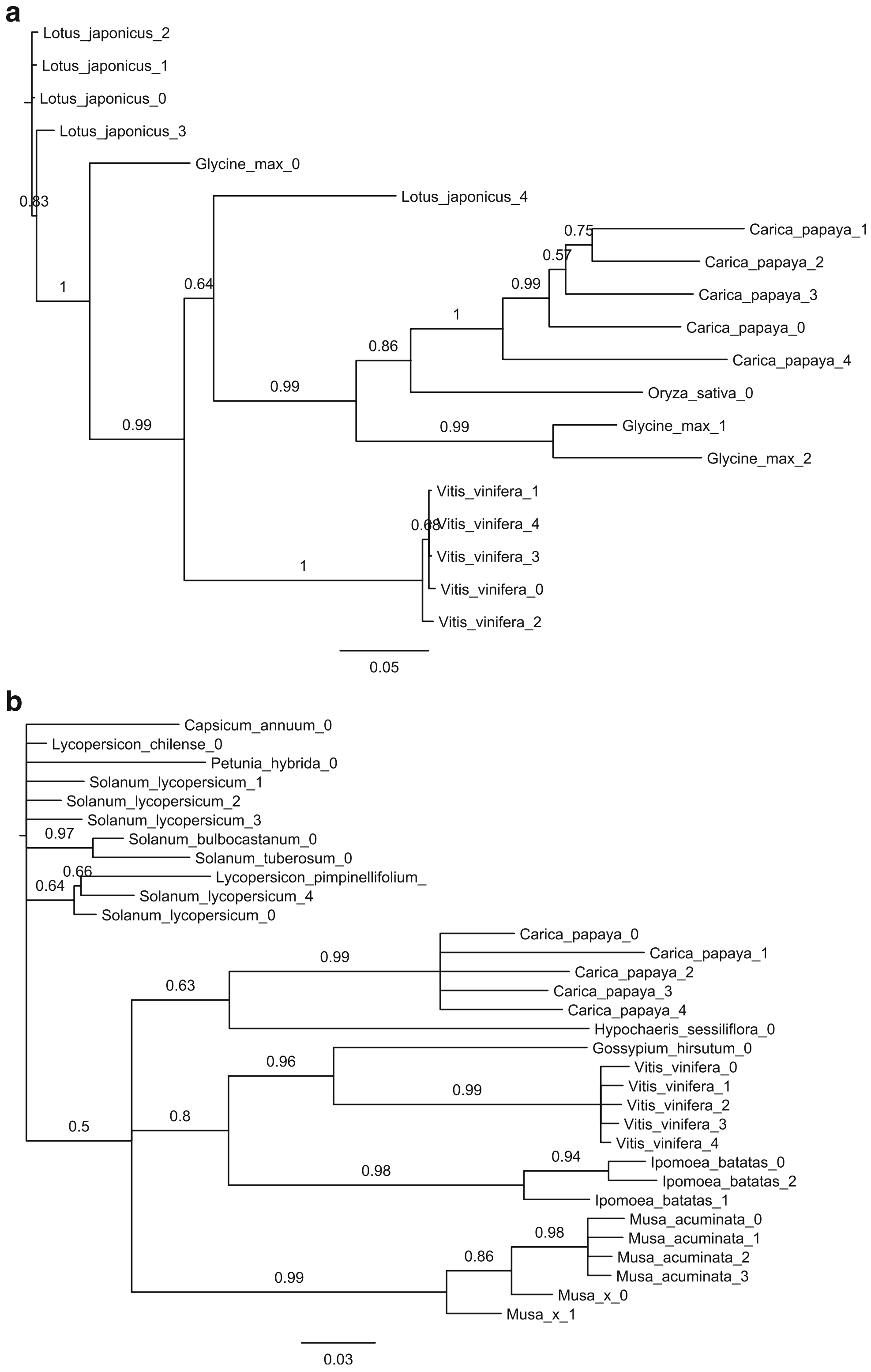


Fig. 1 Phylogenetic analysis of plant genome sequences matching the Ty3-Gypsy retrotransposon sequence ATGP5A_I in **a** bases 1,200 to 1,500 and **b** bases 3,700 to 4,100. Note that where possible, the five best matches for each species were included in the phylogenetic analysis. Also, the edge labels here correspond to the confidence in the edges

of DNA, minisatellites and satellites were the most represented in papaya, constituting 0.68% and 0.43% of the total genome size, respectively. However, microsatellites constitute the class with the highest number of tandem repeat copies, including dinucleotides with ~180,000 units, TTC/AAG, AAT/TTA and their multimeric variants (with up to 969.3 repeats in a single locus) and pentanucleotides as the most common repeats. Several stretches of A/T, AT/TA, TAC/ATG, AGA/TCT and ATT/TAA motifs are also very common in papaya. Significantly, we detected the plant telomeric motif $(TTTAGGG)_n$ 125 times in four different loci. Except for the presence of the vertebrate telomeric $(TTAGGG)_n$ motif in a small copy number (5), and their variants (TTAGGGC, TTAGGGG, TTAGGGT-repeated ~100 times) no other telomere-associated or centromeric motifs were detected. Minisatellites, especially those ranging from 9 to 30 bp, constituted the second most frequent tandem repeat type. Interestingly, for this class we also found the maximum number of loci and sequence variants (Fig. 3). Finally, 2,866 variants met the requirements to be considered putative satellite DNA sequences. Up to 70% of them ranged from 101 to 300 bp in length.

A non-redundant dataset was constructed including 23,041 repeat families. For a small fraction of these (1,790), a hit was found in the *Arabidopsis* genome. The annotated function was classified based on the hit in the MIPS functional catalogue database (http://mips.gsf.de).

Fig. 2 Phylogenetic analysis of plant genome sequences matching the Ty1-Copia retrotransposon sequences **a** RN107_I bases 3,400 to 5,000 and **b** FRSgTERT00100296 bases 1,400 to 1,800. Note that where possible, the five best matches for each species were included in the phylogenetic analysis. Also, the edge labels here correspond to the confidence in the edges



The vast majority of matches corresponded to unclassified proteins, mainly including TE-like sequences, pseudogenes, and DNA-binding factors. The remaining sequences had no match and should be regarded as papaya-specific.

High Copy Number Genes

While most of the genes in the papaya genome have a low copy number (based on a BLAST comparison with the genome, see "Methods"), a significant fraction of the genes

Table 3 Tandem repeats in the papaya genome

	Total number of blocks	Total number of copies	Variants	Total length/genome %
Microsatellites (1–6 bp)				
Mononucleotide	1,165	39,236	4	39,236
Dinucleotide	9,879	182,121	11	361,834
Trinucleotide	2,504	43,544.2	56	130,208
Tetranucleotide	835	9,338.7	75	37,342
Pentanucleotide	1,822	16,933.9	134	85,150
Hexanucleotide	1,448	11,914.4	551	71,255
Total	17,653	303,088.2	831	725,025/0.19%
Minisatellites (7–100 bp)				
7–30 bp	28,500	81,612.9	23,660	1,430,645
31–50 bp	4,767	14,126.7	4,060	501,660
51–70 bp	1,686	4,292.6	1,663	245,130
71–100 bp	1,888	4,423.5	1,844	364,977
Total	35,731	98,228.3	31,227	2,542,412/0.68%
Satellites (>100 bp)				
101–200 bp	1,712	3,864.4	1,704	539,166
201–300 bp	473	1,126.9	471	271,170
301–400 bp	536	1,715.9	536	566,950
>400 bp	145	430.3	145	198,947
Total	2,866	7,137.5	2,856	1,576,233/0.43%
Grand total	57,360	414,681.4	34,914	4,843,670

(>2%, representing 3% of the papaya genome) are present in a large number of copies (>20) as can be seen in Fig. 4. Many of the most abundant genes are similar to those found in TEs (gag-pol polyprotein, retroelement integrase) and it is therefore not surprising to see them in so many copies. Some of the genes, however, represent non-TE related proteins providing interesting insights into papaya biology (Table 4). Comparison of papaya genes at the protein level (see "Methods") revealed a similar but slightly different set of genes that are part of large families in the papaya proteome (Table 5).

Discussion

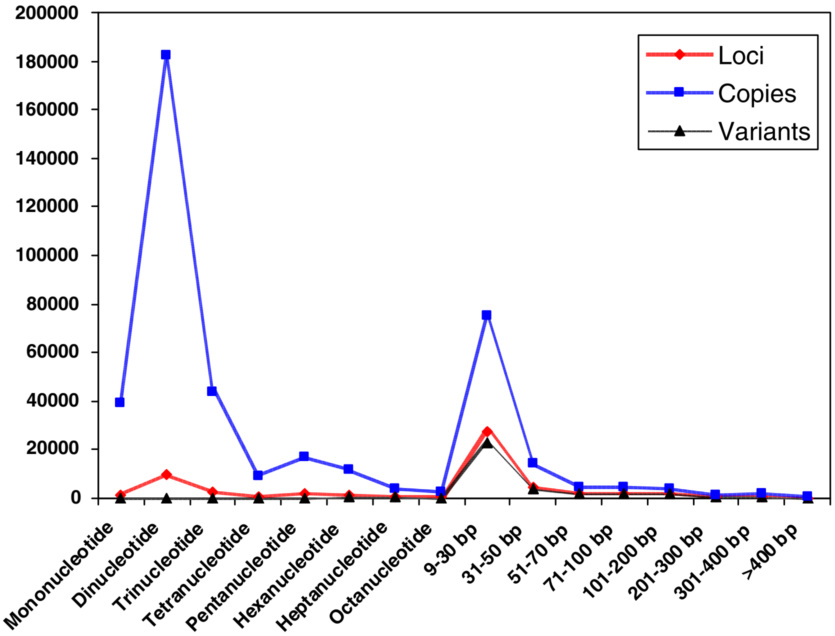
The papaya female genome, as a conservative estimate (due to the difficulty of assembling repeat elements in the whole genome shotgun sequencing approach), has a large fraction (~56%) composed of repetitive sequences. A lion's share of these repetitive sequences is taken up by transposable elements (52%). Since large genomes tend to have higher TE content, the proportion in the papaya genome is somewhat as expected in comparison to the much smaller Arabidopsis genome [2] (14% TE content) and the much larger maize genome [28] (58% TE content). However, in contrast to the larger rice genome [17] (35% TE content), papaya is relatively TE rich. While being roughly three times the size of the Arabidopsis genome, the papaya genome has similar TE content to the maize genome that is nearly twenty times the size of Arabidopsis. The relatively high TE content of the papaya genome also agrees with the observation that while its genome is three times the size of the Arabidopsis genome its gene repertoire is actually smaller [30].

The papaya genome and the rice genome share some similarities in their transposon content. For example, a large fraction of matches to TEs in the papaya genome are to known elements in the rice genome (Table 6). In addition, in contrast to a prior analysis [21], our analysis suggests that the ratio of *Ty3-gypsy* to *Ty1-copia* elements in the papaya genome is closer to the 2:1 ratio[1] of the rice genome than to the 1:1 of *Arabidopsis* and maize genomes [28].

Phylogenetic analysis of TE sequences in the papaya genome reveals a familiar pattern of these sequences tending to cluster with each other and being distinct from homologous sequences from other genomes. An interesting feature seen in these phylogenies is that the within species divergence in two Ty1-Copia elements is greater in papaya than in the grape genome while the opposite seems to be true for a Ty3-Gypsy element. Also, the Ty3-Gypsy element is more conserved over a larger evolutionary time scale, from Arabidopsis to rice to papaya, than for the Ty1-Copia elements. Another intriguing aspect is the tendency of sequences from the rice genome to cluster with papaya sequences, despite their large divergence on the species tree. In our analysis, we observed that papaya sequences

[1] We use the estimate from Table 2 as a large fraction of the Retrotransposon matches in Table 1 have not been classified.

Fig. 3 Distribution of tandem repeats in papaya



tend to have higher substitution rates and this could be an explanation for this discrepancy.

Another important part, constituting ~2%, of the papaya genome, is composed of tandem repeats. Relative to other plant genomes (reviewed in [25]), tandem repeats in papaya appear to be under-represented. However, since repeat regions, especially those that are long and highly conserved, are particularly difficult to handle in sequence assembly, this percentage could be considerably underestimated.

Tandem repeats are more-or-less randomly distributed in the papaya genome, with their number positively correlated with supercontig length. However, no correlation was found between tandem repeat number and gene density (Fig. 5). Since tandem repeats are known to be the main component of constitutive heterochromatin [12], this fact may suggest that the draft genome of papaya lacks assembled contigs of heterochromatin, although DAPI staining experiments suggest that papaya genome is also largely euchromatic [30].

Based on genomic library screenings some repetitive motifs have been found to be abundant in plants. We found that the most common dinucleotides in papaya are $(TA/AT)_n$ and $(AG/TC)_n$ along with long A/T stretches. This agrees with the results of Macas et al. [25] and Lagercrantz et al. [20] who found that $(AA/TT)_n$ and $(AG/TC)_n$ are the

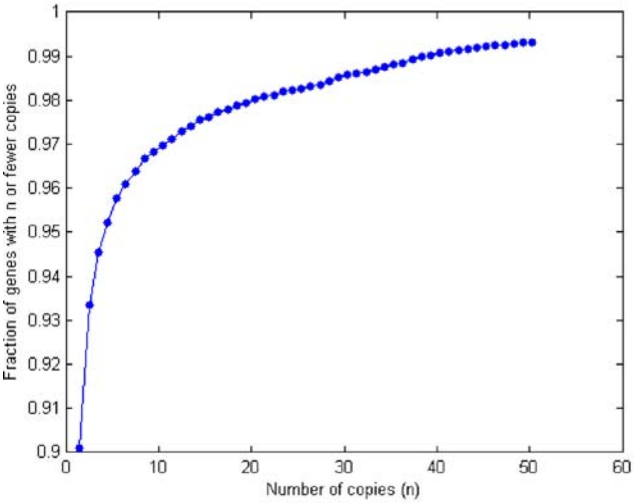


Fig. 4 Distribution of copy numbers for papaya genes (for $1 \leq n \leq 50$)

Table 4 High copy number transcripts in the papaya genome

Transcript	Copy number	Similar to
evm.model.supercontig_185.9	112	MADS box transcription factor
evm.model.supercontig_219.4	100	Zinc finger
evm.model.supercontig_30.166	91	MADS box transcription factor
evm.model.supercontig_3040.1	70	Protein kinase
evm.model.supercontig_2.67	30	NADH-plastoquinone oxidoreductase subunit 2

Note that the numbers reported here are based on independent blast searches and complement the study based on gene models in Ming et al. [30]. In particular, the matches to the two MADS box genes overlap substantially and should not be interpreted as suggesting that there are 112+91 MADS box genes in the papaya genome (for a more authoritative number see Ming et al. [30])

Table 5 High copy number proteins in the papaya genome

Transcript	Copy number	Similar to
evm.model.supercontig_2.153	>500	Topoisomerase I
evm.model.supercontig_13.121	430	Serine/threonine phosphatase
evm.model.supercontig_232.8	416	Guanine nucleotide-exchange protein
evm.model.supercontig_1.174	316	Pentatricopeptide repeat-containing protein
evm.model.supercontig_224.2	137	Salt-inducible protein

most common dinucleotides in plants. Trinucleotides $(TTC/AAG)_n$, $(AAT/TTA)_n$ and $(TAA/ATT)_n$ are also abundant in the papaya genome. $(TAA/ATT)_n$ is also predominant in wheat [46], tomato [45], soybean [1] and *Arabidopsis* [24]. In contrast, $(AAC/TTG)_n$ and $(ACC/TGG)_n$ that account for 84.5% of eggplant microsatellites [34], seem to be under-represented in papaya with only a few copies.

We found a moderate number of copies of the plant telomeric motif $(TTTAGGG)_n$ indicating that papaya telomeres belong to the *Arabidopsis* type, which is not surprising since they are in the same order (*Brassicales*). Nonetheless, we also found the vertebrate type telomeric motif $(TTAGGG)_n$ and some variants several times. The presence of such derived vertebrate-telomeric motifs has also been reported in lily plants [9].

For longer tandem repeats, inferring relationships with other groups of plants is more difficult since these sequences are normally specific to a related group of species [33, 42] and also undergo rapid evolutionary changes [29]. In this context, it is not surprising that BLAST hits with the *Arabidopsis* genome were found only for a small portion of repeats. In these cases, they showed homology to DNA binding sequences, pseudogenes, or TE-like DNA sequences, suggesting their possible role in gene regulation/inactivation and their probable origin as TEs [26, 27, 37]. These findings agree with some recent studies that indicate that tandem repeats could have an important role in gene regulation processes [48], speciation [16], or proper chromosomal packing in mitosis and meiosis [8] and contrast with earlier studies that denied any function for these sequences, considering them to be junk [35], or parasitic elements [36].

Interestingly, we detected a bias in the distribution of repeat-unit sizes in papaya tandem repeats. It appears that sequences between 9 and 50 bp account for a high number of copies as well as for the maximum number of variants and loci (Fig. 3). It could indicate that tandem repeat units in this range are preferred in the papaya genome. Some authors have argued that structural features such as monomer length, AT content, short sequence motifs or secondary and tertiary structures may be important factors for tandem repeat preservation and evolution [13, 39, 49]. It has been proposed that these structural constraints could be important for tight packing of DNA and proteins in heterochromatin, and are consequently under selective pressure [49] and this could be an important area for future studies.

Finally, despite the fact that the papaya genome contains fewer genes than the *Arabidopsis* genome [30], several gene families have strikingly higher copy number in papaya than *Arabidopsis*, particularly in families associated with tree and fruit development. The set of high copy number genes identified here are therefore interesting targets for further study and characterization to reveal their functions in the papaya genome. Note that the analysis here is complementary to the analysis in Ming et al. [30] where a comparative analysis with other sequenced plant genomes was used to find gene families under strong selection.

Methods

Transposable Elements Analysis

We used a combination of homology-based and de novo methods to identify signatures of transposable elements (TEs) in the papaya genome. Because there are many known families of TEs in plants, homology-based methods should be highly effective in identifying and annotating them. We used RepeatMasker (http://www.repeatmasker.org) (which combines BLAST searches with an array of heuristics to

Table 6 Summary of matches to TEs in various TIGR plant repeat databases

Plant	Class I (retrotransposons)		Class II (transposons)	
	Estimated count	Length covered (in kbp)	Estimated count	Length covered (in kbp)
Arabidopsis	556	128	12	1.6
Brassica	696	171	14	1.6
Glycine	444	165	1	0.05
Hordeum	292	66	956	59
Lycopersicon	2,744	1,200	1	0.37
Lotus	127	29	0	0
Medicago	323	39	6	0.94
Oryza	13,719	5,300	7,542	552
Solanum	550	125	321	65
Sorghum	796	187	17	2.1
Triticum	1,281	257	564	43
Zea	7,867	1,600	129	17
Total	29,395	9,300	9,563	684

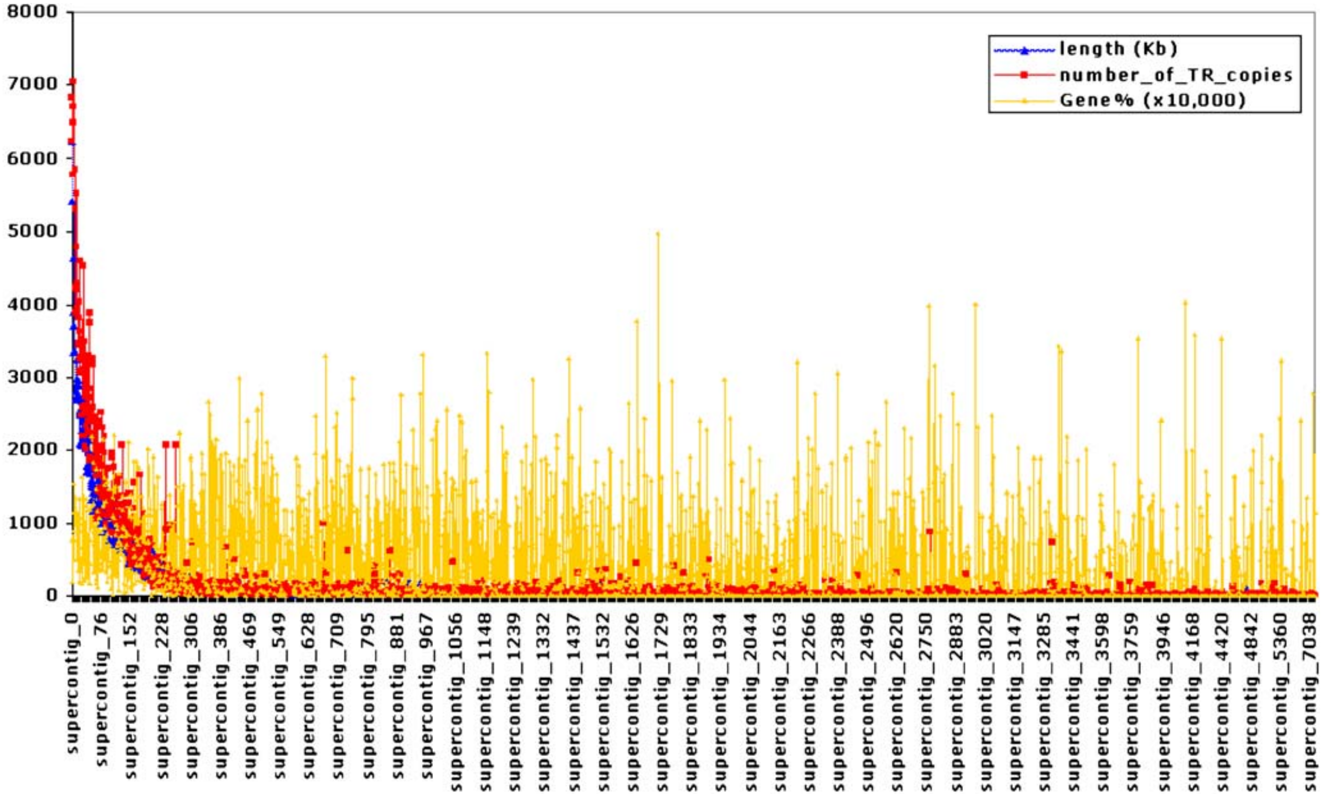


Fig. 5 Distribution of tandem repeat copy number, supercontig length and gene density for papaya scaffolds

organize the matches) in combination with a custom-built library of plant repeat elements for our initial classification of TEs (Tables 2 and 3). The customized library was generated by combining plant repeats from Repbase [18] and plant repeat databases from TIGR (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats). Repeat elements identified as ribosomal RNA sequences in the TIGR databases (matching ~3% of the papaya genome) were excluded from our repeat library, leaving a database of 76,924 repeat sequences that were used to search the papaya genome.

Homology-based methods are limited to finding elements that have not diverged too greatly from known repeats. Because databases of known TEs are necessarily incomplete, we used additional de novo methods to search for repeat elements in papaya contigs. For this purpose, we applied two recently developed repeat-finding tools, PILER [41] and RepeatScout [11] to the complete set of contigs from the papaya genome. Both these tools are fast (RepeatScout ran in less than 4 h) and can process large genomes on a standard desktop Linux computer. In all, PILER was able to find 428 repeat families while RepeatScout found 6,596 repeat sequences. Note that the output from RepeatScout is not grouped into families and hence the repeat sequences that it finds tend to be redundant.

The repeat families obtained from PILER and RepeatScout were annotated using a combination of manual curation (786 repeat families, N. Jiang, personal communication) and automated analysis. For the automated annotation, the combined dataset from PILER and RepeatScout was made non-redundant, using CD-HIT [23] at the 90% similarity level, leaving behind 6,240 repeat families. As a post-processing step, we selected only those families which have at least ten good (E-value$<10^{-20}$) BLAST matches to papaya contigs. The resulting dataset contains 2,198 repeat families in the papaya genome (84 found by PILER and 2,114 found by RepeatScout). BLAST searches against NR and PTREP (http://wheat.pw.usda.gov/ITMI/Repeats) were then used to identify repeat families matching genes associated with transposons and retrotransposons. This procedure discovered an additional 103 repeat families that could be annotated as retrotransposons. The combined database of 889 annotated papaya-specific TE sequences was used in addition to the database of known repeats to annotate the papaya genome (Table 1). The remaining, unannotated repeat families (1,455 sequences with no matches to known genes) were then used to estimate the additional repeat content of the genome (the "Unannotated" class in Table 1).

Phylogenetic Analysis of TE Sequences

From the set of known plant TE sequences, we identified three sequences with many good matches in the papaya genome for further phylogenetic analysis. Two of these sequences (RN107_I and FRSgTERT00100296) correspond to Ty1-Copia retrotransposons while the third (ATGP5A_I) corresponds to a Ty3-Gypsy restrotransposon element. BLAST searches against the papaya genome and the NR DNA database (NCBI, January 2008) was used to find sequences with good homology (E-value$<10^{-20}$). These searches helped identify the conserved regions of these sequences which were then used to generate CLUSTALW [22] multiple alignments for these sequences (up to five sequences for each species). Phylogenetic tree reconstruction was then performed using MrBayes [43] to generate an ensemble of 500 trees (GTR model with gamma distributed rate variation among sites) and derive the consensus tree. Interestingly, in all cases, the TE sequences only matched regions in other plant genomes.

Tandem Repeats Detection

The papaya whole genome sequence was explored for tandem repeats by using the Tandem Repeats Finder software [5]. Repeat units between 1 and 2,000 bp were analyzed, and only repeats arrayed in tandems ≥25 bp were considered (a complete catalogue of SSRs is described in Wang et al. this issue). Repeats were classified into micro- (1–6 bp), mini- (7–100 bp) and satellite (>100 bp) tandemly-arrayed sequences. A non-redundant set of sequences was constructed using the program cd-hit-est, as implemented in the package CD-HIT [23], at the 85% similarity level. For annotations, the non-redundant sequences were BLASTed with the *Arabidopsis* TAIR 7 release [40] and the hits classified according to the MIPS functional catalogue database (http://mips.gsf.de). Perl scripts were written to automate the process.

High Copy Number Genes Detection

The set of annotated genes (including introns and exons) in the papaya genome were BLASTed against the whole genome sequence to find significant matches (E-value$<10^{-20}$). Similar searches were also conducted using the predicted protein sequences. In addition, the papaya genes were annotated by BLASTing against the NR protein database (NCBI, January 2008) and transferring annotations from the best match if it was a significant match (E-value$<10^{-20}$).

Data Access and Retrieval

The sequences and annotations in the CPR-DB database are available via FTP downloads at ftp://ftp.cbcb.umd.edu/pub/data/CPR-DB. The sets of novel TE sequence in papaya (annotated and un-annotated) are presented as multi-fasta files in a format convenient for use with RepeatMasker. For tandem repeats, redundant and non-redundant databases as well as a consensus sequence list are available in multi-fasta files. A file including annotations is also provided. High-copy number papaya transcripts and protein sequences are also available as annotated multi-fasta files. Further details can be found in the README file accompanying the database.

Acknowledgements We thank Ning Jiang for suggestions and discussion on construction and characterization of the papaya repeat database. We appreciate financial support from the U.S. National Institutes of Health (R01-GM083873 to S.S.), the U.S. National Science Foundation (DBI-0553417 to R.M. and A.H.P.), the U. Hawaii and U.S. Department of Defense (W81XWH0520013 to M.A).

References

1. Akkaya MS, Shoemaker RC, Specht JE, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci 35:1439–1445
2. Arabidopsis Genome Initiative (2001) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815 doi:10.1038/35048692
3. Bennetzen JL (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115:29–36 doi:10.1023/A:1016015913350
4. Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot (Lond) 95:127–132 doi:10.1093/aob/mci008
5. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580 doi:10.1093/nar/27.2.573
6. Camacho JP, Sharbel TF, Beukeboom LW (2000) B-chromosome evolution. Philos Trans R Soc Lond B Biol Sci 355:163–178 doi:10.1098/rstb.2000.0556
7. Cheng XD, Ling HQ (2006) Non-LTR retrotransposons: LINEs and SINEs in plant genome. Yichuan 28:731–736
8. Csink AK, Henikoff S (1998) Something from nothing: the evolution and utility of satellite repeats. Trends Genet 14:200–204 doi:10.1016/S0168-9525(98)01444-9
9. de la Herrán R, Cuñado N, Navajas-Pérez N, Santos JL, Ruiz Rejón C, Garrido-Ramos MA et al (2005) The controversial telomeres of lily plants. Cytogenet Genome Res 109:144–147 doi:10.1159/000082393
10. de Ridder C, Kourie DG, Watson BW (2006) FireμSat: meeting the challenge of detecting microsatellites in DNA. Proc SAICSIT 2006:247–256 doi:10.1145/1216262.1216289
11. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21:i152–i158 doi:10.1093/bioinformatics/bti1003
12. Elder JR, Turner BJ (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. Q Rev Biol 70:297–320 doi:10.1086/419073
13. Fitzgerald DJ, Dryden GL, Bronson EC, Williams JS, Anderson JN (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. J Biol Chem 269:21303–21314
14. Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and proportion of repeated nucleotide-sequence DNA in plants. Biochem Genet 12:257–269 doi:10.1007/BF00485947
15. Hatch FT, Mazrimas JA (1974) Fractionation and characterisation of satellite DNAs of the kangaroo rat (*Dipodomys ordii*). Nucleic Acids Res 1:559–575 doi:10.1093/nar/1.4.559

16. Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293:1098–1102 doi:10.1126/science.1062939
17. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800
18. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467 doi:10.1159/000084979
19. Kubis SE, Schmidt T, Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Ann Bot (Lond) 82:45–55 doi:10.1006/anbo.1998.0779
20. Lagercrantz U, Ellegren H, Andersson L (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res 21:1111–1115 doi:10.1093/nar/21.5.1111
21. Lai CW, Yu Q, Hou S, Skelton RL, Jones MR, Lewis KL et al (2006) Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome. Mol Genet Genomics 276(1):1–12 doi:10.1007/s00438-006-0122-z
22. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H et al (2007) ClustalW and ClustalX version 2. Bioinformatics 23(21):2947–2948 doi:10.1093/bioinformatics/btm404
23. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659 doi:10.1093/bioinformatics/btl158
24. Loridon K, Cournoyer B, Goubely C, Depeiges A, Picard G (1998) Length polymorphism and allele structure of trinucleotide microsatellites in natural accessions of *Arabidopsis thaliana*. Theor Appl Genet 97:591–604 doi:10.1007/s001220050935
25. Macas J, Mészáros T, Nouzová M (2002) PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18:28–35 doi:10.1093/bioinformatics/18.1.28
26. McCombie WR et al (2000) The complete sequence of a heterochromatic island from a higher eukaryote. Cell 100:377–386 doi:10.1016/S0092-8674(00)80673-X
27. Meagher TR, Vassiliadis C (2005) Phenotypic impacts of repetitive DNA in flowering plants. New Phytol 168:71–80 doi:10.1111/j.1469-8137.2005.01527.x
28. Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y et al (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci U S A 101:14349–14354 doi:10.1073/pnas.0406163101
29. Miklos GL (1985) Localited highly repetitive DNA sequences in vertebrate and invertebrate genomes. In: McIntryre JR (ed) Molecular evolutionary genetics. Plenum, New York, pp 231–241
30. Ming R et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996 doi:10.1038/nature06856
31. Murray MG, Peters DL, Thompson WF (1981) Ancient repeated sequences in the pea and mung bean genomes and implications for genome evolution. J Mol Evol 17:31–42 doi:10.1007/BF01792422
32. Navajas-Pérez R, Rubio-Escudero C, Aznarte JL, Ruiz Rejón M, Garrido-Ramos MA (2007) SatDNA Analyzer: a computing tool for satellite-DNA evolutionary analysis. Bioinformatics 23:767–768 doi:10.1093/bioinformatics/btm005
33. Navajas-Pérez R, Schwarzacher T, de la Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA (2006) The origin and evolution of the variability in a Y-specific satellite-DNA of *Rumex acetosa* and its relatives. Gene 368:61–71 doi:10.1016/j.gene.2005.10.013
34. Nunome T, Suwabe K, Ohyama A, Fukuoka H (2003) Characterization of trinucleotide microsatellites in eggplant. Breed Sci 53:77–83 doi:10.1270/jsbbs.53.77
35. Ohno S (1972) So much “junk” DNA in our genome. Brookhaven Symp Biol 23:366–370
36. Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607 doi:10.1038/284604a0
37. Pelissier T, Tutois S, Tourmente S, Deragon JM, Picard G (1996) DNA regions flanking the major *Arabidopsis thaliana* are principally enriched in Athila retroelement sequences. Genetica 97:141–151 doi:10.1007/BF00054621
38. Petitpierre E, Juan C, Pons J, Plohl M, Ugarković D (1995) Satellite DNA and constitutive heterochromatin in tenebrionid beetles. In: Brandham PE, Bennett MD (eds) Kew chromosome conference IV. Royal Botanic Gardens, London, pp 351–362
39. Plohl M, Mestrovic N, Bruvo B, Ugarkovic D (1998) Similarity of structural features and evolution of satellite DNAs from *Palorus subdepressus* (Coleoptera) and related species. J Mol Evol 46:234–239 doi:10.1007/PL00006298
40. Poole RL (2007) The TAIR Database. Methods Mol Biol 406:179–212 doi:10.1007/978-1-59745-535-0_8
41. Price AL, Jones NC, Pevzner PA (2005) *De novo* identification of repeat families in large genomes. Bioinformatics 21:351–358 doi:10.1093/bioinformatics/bti1018
42. Rajagopal J, Das S, Khurana DK, Srivastava PS, Lakshmikumaran M (1999) Molecular characterization and distribution of a 145-bp tandem repeat family in the genus *Populus*. Genome 42:909–918 doi:10.1139/gen-42-5-909
43. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574 doi:10.1093/bioinformatics/btg180
44. Schmidt AL, Anderson LM (2006) Repetitive DNA elements as mediators of genomic change in response to environmental cues. Biol Rev Camb Philos Soc 81:531–543 doi:10.1017/S146479310600710X
45. Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. Theor Appl Genet 97:264–272 doi:10.1007/s001220050409
46. Song QJ, Fickus EW, Cregan PB (2002) Characterization of trinucleotide SSR motifs in wheat. Theor Appl Genet 104:286–293 doi:10.1007/s001220100698
47. Thomas CA Jr (1971) The genetic organization of chromosomes. Annu Rev Genet 5:237–256 doi:10.1146/annurev.ge.05.120171.001321
48. Thornburg BG, Gotea V, Makałowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365:104–110 doi:10.1016/j.gene.2005.09.036
49. Ugarković D, Plohl M (2002) Variation in satellite DNA profiles, causes and effects. EMBO J 21:5955–5959 doi:10.1093/emboj/cdf612
50. Volfovsky N, Haas BJ, Salzberg SL (2001) A clustering method for repeat analysis in DNA sequences. Genome Biol 2(8): research0027.1–research0027.11
51. Wicker T, Matthews DE, Keller B (2002) TREP: a database for Triticeae repetitive elements. Trends Plant Sci 7:561–562 doi:10.1016/S1360-1385(02)02372-5